Variable Selection to Determine Majors of Student using K-Nearest Neighbor and Naïve Bayes Classifier Algorithm

Journal of Physics: Conference Series **1363** (2019) 012057 doi:10.1088/1742-6596/1363/1/012057

Variable Selection to Determine Majors of Student using K-Nearest Neighbor and Naïve Bayes Classifier Algorithm

Mustakim[1,3*], Reysa Hastarimasuci[1,3], Petir Papilo[1], Zarkasih[2], Zaitun[2], Alwis Nazir[1],

[1]Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, 28293, Indonesia.
[2]Faculty of Education and Teching, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, 28293, Indonesia.
[3]Puzzle Research Data Technology, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, 28293, Indonesia.

*mustakim@uin-suska.ac.id

Abstract. Appropriate student's major placement in high school can help students to better improve their academic achievement. There are many variables which must be considered to determine the student's majors, such as: Gender, Interests, Intelligence Quotient (IQ); Four subjects in Junior High School (JHS), average junior high school grades, matriculation score of four subjects, and average rate of matriculation. The number of variables used in the selection, causes some weaknesses among them, such as the complexity of variable, the inefficiency of variable and the existence of some variables which is only as an addition without having a significant contribution. This study aims to reduce the number of these variables so it will become easier to analyze and to be applied. The Process of reduction was done by combining experiments with Predefined attributes. A total of ten combinations were attempted using K-Nearest Neighbor (K-NN) and Naïve Bayes Classifier (NBC) which then was measured by Confusion Matrix accuracy. The experimental result showed that the combination of variables which produce the best accuracy were the 9th and 10th experiment with variable matriculation, interest, and IQ, and an accuracy of 96.77% from K-NN also 98.38% from NBC. By combining both algorithms, 99.87% of maximum accuracy was obtained from those three variables. New information which can be extracted from this research is that there are only three important variables to determine major placement in Senior High School, Average Scores of Matriculation, Interests and IQ followed by four supporting variables such as the scores of Mathematic, Physics, English and Economics in Matriculation.

1. Introduction

Abundant data opens the opportunities to implement data mining techniques for better management of education in the implementation of learning with the help of intelligent computers effectively [1]. Currently, Islamic of Senior High School 2 Model Pekanbaru has 2 majors based on curriculum 2013, namely Mathematics of Natural Sciences (MNS) and Social Sciences major (SS). This major placement aims to provide direction for learners to focus more on developing their potential and interest. At the beginning of school entry, new students are required to attend the matriculation class for a month. Then the scores obtained from the matriculation exams will be taken into consideration and calculated with Junior High School Grades. In addition to the matriculation scores and junior high school grades,

students are also given a questionnaire of interest, they were asked to fill personal information and majors of interest. Furthermore, the teachers of Counseling Guidance (CG) and Curriculum Section will decide the right majors for the students.

The majors placement Process has several problems such as (1) The abundant data will spend a lot of time, energy and needs extra-precision; (2) There are some students and parents who neglect the accuracy of data Processing based on academic potential or score selection which has been recapitulate; (3) Determination of majors which follows the wish of students or parents without regard to their academic grades causes students to experience problems in the future. In addition there are many variables used in determining the majors, some of them are gender; interest; Intelligence Quotient (IQ); Four grades of subjects at Junior High School (JHS) such as Mathematics, Science, Social Studies, and English; Average grades of junior high school subjects; Four grades of matriculation subjects namely Mathematics, Physics, English, and Economics; As well as the average score of matriculation.

The number of variables used as a reference caused problems for both the students and the school. The problem arose because of the high complexity of variables decreased the accuracy of the actual choice, the inefficiency of variable lead to difficulty in identifying the used comparison and the existence of some variables as a supporter without having a significant contribution. Therefore, this study analyzed and reduced the thirteen variables into several variables to minimize these three problems. The expected output is the determinant variable reduction with Data Mining technique so it can represent the thirteen variables used with high accuracy of a combination series used.

Data mining consists of various techniques which can be used to make prediction and classifications, where this technique estimates the possibility that will occur in the future by looking at some information and existing data patterns [2], [3]. Data Mining is a technique which implements several algorithms in solving complex problem [4], either stand-alone, comparing algorithms or combination for extracting knowledge [5]. The most common technique used in problem solving is classification [6]. K-Nearest Neighbor (K-NN) algorithm is one of classification techniques included in the Top Ten Algorithm [7] also Naïve Bayes Classifier (NBC) which is often used in probabilistic [8], [9].

KNN is an algorithm which works by calculating the shortest distance between data attributes [10], [11], it has a high-performance computing [12], [13], a simple algorithm for large data in characteristic [14], [15] also good in terms of accuracy and Performance [12], [13]. While NBC algorithm is prediction technique based on simple probabilistic on Bayes theorem with strong independence assumption (independent) [8], [16]. The advantage of NBC is it's almost similar with KNN in its simple algorithm, it needs small training data and reliable against irrelevant attributes [9], [16]. The study conducted by Yusra, et al compares K-NN and NBC to classify the students' final task majoring in informatics engineering with the result of each accuracy is 86% with the value of k=7 for K-NN and 87% for NBC [17]. Also the combination between K-NN and NBC was done by Putri, et al to classify the employment status on the number of residents in the labor force in Demak for 2012. The combination result had a value of accuracy 96.06% for K-NN and 94.09% for NBC [9].

In this case K-NN and NBC will classify the dataset from each algorithm, the result of variable reduction in K-NN will be reclassified using NBC, and vice versa until it obtains the efficiency value of variable reduction. The essence to be generated is how much each attribute has affects in problem solving, eliminate the attributes which have no effect on the major placement and to find the best attribute combination from the two used algorithms.

2. Material and Method

2.1. Data Mining

Data mining is a Process which use statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from large databases [18]. Data mining also known as Knowledge Discovery in Database (KDD) is used to extract models that describe important class of data [19]. Before doing the data mining Process, it needs some stages called preprocessing [20]. This is because preprocessing techniques have a significant impact on performance of machine learning algorithms. Good database design and good analysis can reduce the problem of missing data through Processing [21].

2.2. K-Nearest Neighbor

The K-NN algorithm was first introduced by Fix and Hodges in 1951 and 1952 [10]–[12]. This algorithm is also one of the lazy learning techniques. K-NN is done by searching k-group objects in the closest training data (similar) to objects in new data or data testing [13], [15]. K-NN is included in data mining method of classification based on learning by analogy. The training data sample has a numerical dimension attribute. Each sample is a point in the n-dimensional space. All training samples is stored in n-dimensional space. When testing the data, K-NN will find the value of k closest to data testing. The proximity is defined in terms of Euclidean distance between two points X = (x1, x2, ..., xn and Y = (y1, y2, ..., yn) [10]–[12], [22].

2.3. Naïve Bayes Classifier

The Naive Bayes Classifier (NBC) algorithm is a statistical classification method based on Bayes theorem [8]. The Naive Bayes classification assumes the existence of particular feature of a class which has nothing to do with the characteristics of other classes. Naïve Bayes has the potential to classify the data because of its simplicity [17]. The equation of Bayes Theorem is [23]:

$$P(H \mid X)=(P(H \mid X).P(H))/(P(X)) \quad (1)$$

Note:

X : Data with unknown class
H : The data hypothesis of X is a specific class
$(H|X)$: The probability of hypothesis H based on X (posteriori Probability)
(H) : Probability of hypothesis H (Prior Probability)
$(X|H)$: The Probability of X based on the conditions of hypothesis H
(X) : Probability of X.

2.4. Confusion Matrix

Confusion matrix is a tool used to evaluate the classification model in estimating the correct or false object. A matrix of Prediction which will be compared to the original input class which contains information of actual value and Predictions on the classification. Confusion Matrix is a classification evaluation model based on data testing and every Predicted data with the right Proportions [24]. The formula to calculate the accuracy [25]:

$$\text{Accuracy}= \frac{\text{number of correct data}}{\text{total data}} \text{ x } 100\% \quad (2)$$

Table 1. Confusion Matrix of 2 Class [26]

Classification	Prediction Class	
	Class = Yes	Class = No
Class = Yes	a (true positive TP)	b (false negative FN)
Class = No	c (false positive FP)	d (true negative TN)

3. Research Methodology

The stage in calculating K-NN and NBC algorithm in this research was done by applying variable simulation, by combining variables in the classification Process using K-NN and NBC algorithm, can be seen in Figure 1.

4. Results and Discussion

The data used in this study was taken from data recapitulation of student majors in Islamic of Senior High School 2 Model Pekanbaru by CG Teachers and Curriculum Section academic year 2013/2014 to 2016/2017. The number of data collected is 356 records which consist of 227 records for training data and 153 records for testing data. The student majors that have been collected will be used for the classification Process in order to obtain the best features or variables from the simulation Process in the

placement of majors. A total of 13 variables were used in the Process of classification in this study with 2 targets, Mathematics Natural Sciences (MNS) and Social Sciences (SS).

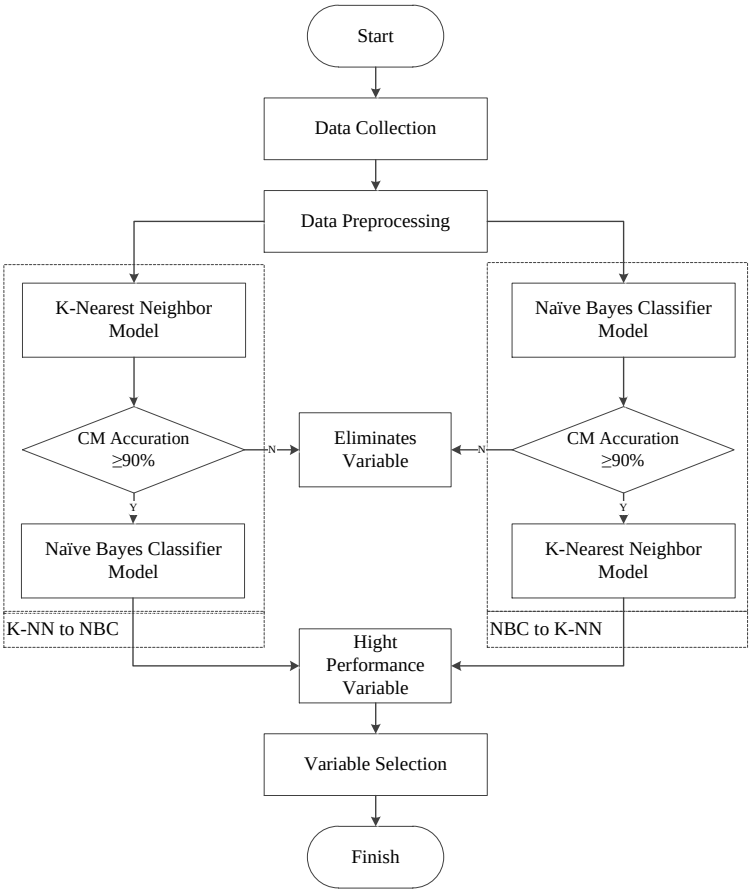

Figure 1. Research Methodology

The data taken is the data value based on the success parameters of Students in majors which have been implemented for four semesters in Senior High School. There are two determined success parameters, high school grades from semesters one to four on compulsory subjects of MNS for science majors ≥ 80 and compulsory subjects SS for SS majors ≥ 77 with graphics value increased for four semesters.

Smoothing technique was done to data generated from schools using binning method by applying equal width distance [25]. It is done to avoid data which contains noise or invalid. At this stage equal width is used to divide the interval of attribute values into n equal width intervals. Then Proceed with transformation by changing equal width distance data into low range = 1, medium = 2, and high = 3.

Table 2. Dataset Students' Grades in Senior High School

Number	Student ID	Gender	IQ	Interest	Matriculation Scores				AVG Matriculation	Junior High School Grades				AVG Grades
					Math	Physic	English	Eco		Math	Science	Social	English	
001	1310A01	M	116	MNS	40	47	52	40	44.75	94.8	89.0	92.4	93.2	92.35
002	1310A02	M	131	MNS	80	68	76	70	73.50	93.8	90.2	94.0	88.8	91.70
003	1310A03	F	109	MNS	84	75	52	70	70.25	90.8	86.2	91.4	90.2	89.65
004	1310A07	F	124	MNS	76	63	56	50	61.25	84.0	82.0	81.2	81.2	82.10
005	1310A08	M	131	MNS	60	60	68	70	64.50	93.4	92.8	89.6	91.6	91.85
006	1310S41	M	109	SS	80	67	60	60	66.75	82.6	80.4	85.0	83.0	82.75
007	1310S42	M	101	SS	20	51	28	50	37.25	80.2	82.6	81.6	86.8	82.80
…	…	…	…	…	…	…	…	…	…	…	…	…	...	...
875	1310S62	P	97	SS	28	30	45	43	36.50	73.5	77.8	78.2	74.4	75.97

4.1. Variable Simulation

Thirteen variables used will be combined and divided into ten simulation stages as follows: (1) All Variables (Gender, IQ, Interest, Matriculation Scores, Junior High School Grades); (2) Junior grades (Mathematics, Science, SS, and English); (3) Matriculation Scores (Mathematics, Physics, English, and Economics); (4) Junior High grades with IQ; (5) Matriculation Scores with IQ; (6) Junior high school grades with interests; (7) Matriculation Scores with interest; (8) Average grades of junior high school with IQ; (9) Average grades of Matriculation with Interests; (10) IQ with Interest.

4.2. KNN and NBC Analysis

The experimental results showed that the minimum accuracy value of K-NN based on confusion matrix is 67.88% in the third simulation, while 2 simulations have accuracy above 90% in the ninth simulation (S-9) and tenth (S-10) by 94.83% and 96.77% respectively.

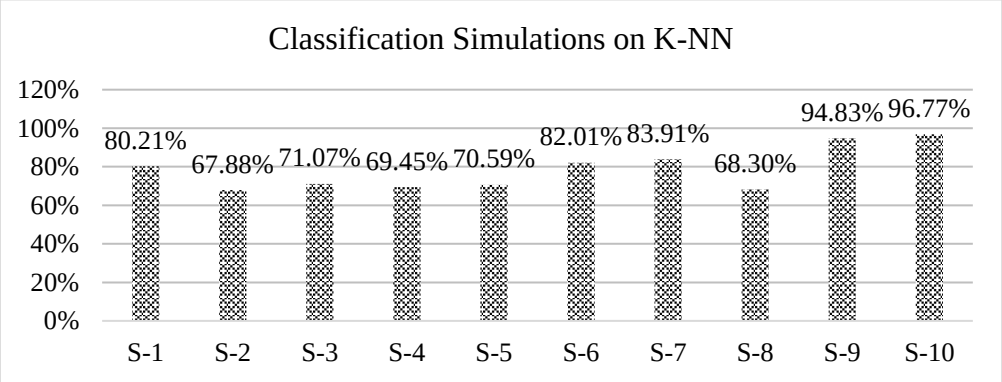

Figure 2. Results of K-NN Accuracy using Confusion Matrix

From figure 2 above it can be seen that S-9 and S10 are the highest simulation. So the best variable on K-NN algorithm to be tested on NBC algorithm is the 9th simulation with attribute of average matriculation and interest, and the 10th simulation with attribute of interest and IQ. For modeling NBC by applying the best variables (S9 and S-10) from the accuracy of K-NN algorithm in previous stages. Based on the classification experiments using NBC algorithm and accuracy testing using confusion matrix, the average matriculation scores, interest and IQ variables obtain 99.87% of accuracy for the three attributes on the NBC algorithm.

4.3. NBC and KNN Analysis

NBC's analysis on K-NN algorithm has some different accuracy results, especially in terms of Confusion Matrix accuracy. Here are the results of NBC accuracy on K-NN.

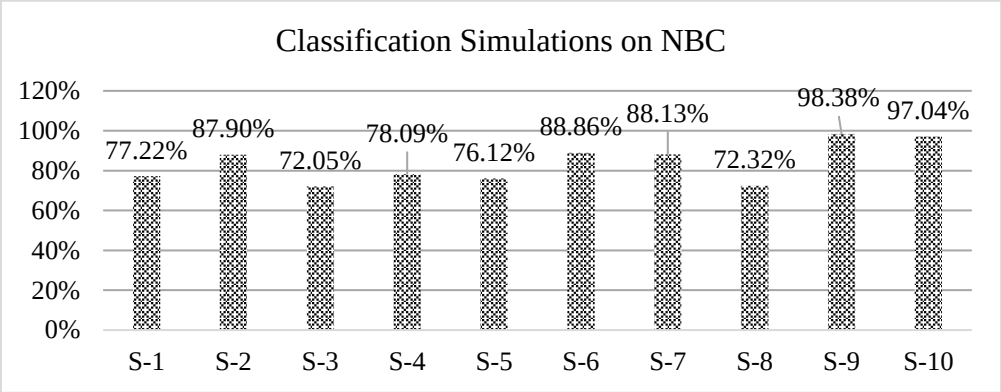

Figure 3. Results of NBC Accuracy using Confusion Matrix

From figure 3 above it can be seen that classification using NBC has better tendency than K-NN, it can be seen from the accuracy of confusion matrix with minimum accuracy value of 72,05% at S-3. Accuracy results also show that S-9 and S-10 have values above 90% as well as in K-NN, which can be interpreted that the three variables namely the average matriculation scores, interest and IQ are important variable in this study. Next, just like previous experiments, the highest accuracy greater than 90% will be tested using K-NN. New combinations of these three attributes result in accuracy of 99.03%.

The overall results of comparison between K-NN and NBC experiments as well as NBC experiments with K-NN can be seen in Figure 4 below:

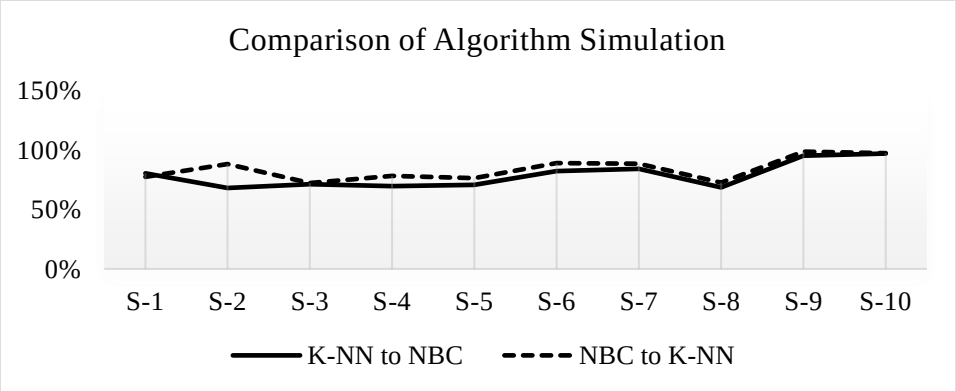

Figure 4. Comparison of Accuracy between K-NN and NBC

While the combination of experiments S-9 and S-10 which consists of three main variables on each of experimental algorithm can be seen in Figure 5 below.

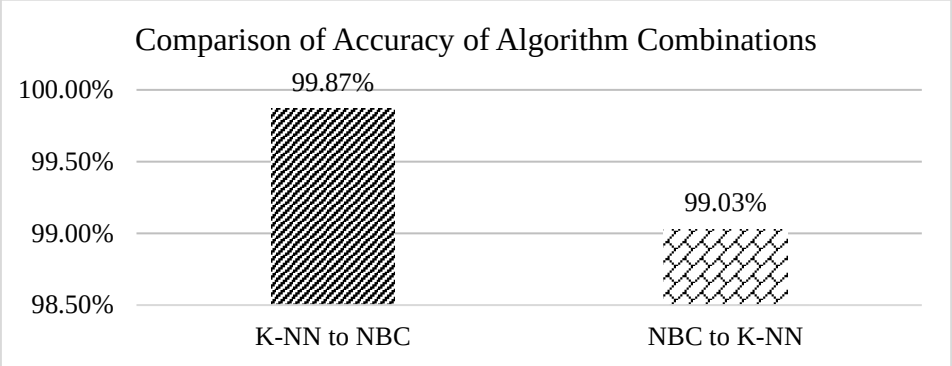

Figure 5. Comparison of Best Combination Accuracy of K-NN and NBC

According to comparative analysis on both experimental simulation and combination algorithm using K-NN and NBC, a new knowledge is formed from a set of variables in the dataset that the result of combining attributes in each simulation shows the three best attributes of all attributes used are the average matriculation scores, interest, and IQ. The average scores of matriculation consist of Mathematics, Physics, English, and Economics are the supporting variables used as school reference in selecting the best majors. Other factors or variables such as Junior High grades consist of Mathematics, Science, Social Studies, and English; Average grades of junior high school subjects; The average value of matriculation and gender cannot be used as a reference in the selection of majors.

5. Conclusion

The calculation of NBC algorithm applied for student major classification in Islamic of Senior High School 2 Pekanbaru Model is able to produce the highest accuracy of 98.38% at S-9, it is higher than K-NN 96.77% at S-10. S-9 and S-10 contain combination of main variable, the average of matriculation scores with interest and interest with IQ. So it can be concluded that to determine the major placement, the most important variables which need to be the main references are the average scores of Matriculation, Interests and IQ. The three variables were also combined to two algorithms, K-NN and NBC. Both algorithms were compared by applying the accuracy greater than 90% which obtained the accuracy of K-NN to NBC by 99.87% and NBC to K-NN equal to 99.03%. The result of this research in general can be stated that there are only 3 main variables and 4 matriculation scores as supporting variables that can be used as reference in the major placement, not thirteen attributes as previously proposed.

Acknowledgment

Acknowledgments to Dean Faculty of Science and Technology (FST) Universitas Islam Negeri (UIN) Sultan Syarif Kasim Riau who has facilitated the research budget for lecturers and students, provides laboratory facilities and space for Research Center for Large Data and thanks to Institutions of Research and Community Service UIN Sultan Syarif Kasim Riau. Also to Tim Puzzle Research Data Technology (Predatech) FST thank you for your cooperation, motivation and persistence in doing research in the field of Data Mining and Big Data.

References

[1] M. A. Nurrohmat, "Aplikasi Pemrediksi Masa Studi dan Predikat Kelulusan Mahasiswa Informatika Universitas Muhammadiyah Surakarta Menggunakan Metode Naive Bayes," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 1, no. 1, pp. 29–34, 2015.

[2] M. Mustakim and E. Saputra, "Aplikasi Prediksi Hasil Tanaman Palawija di Kabupaten Indragiri Hilir Menggunakan Metode Marcov Chains," *J. Sains dan Teknol. Ind.*, vol. 9, no. 2, pp. 50–59, 2014.

[3] J. Archenaa and E. A. M. Anita, "A survey of big data analytics in healthcare and government," *Procedia Comput. Sci.*, vol. 50, pp. 408–413, 2015.

[4] D. Tomar and S. Agarwal, "A survey on Data Mining approaches for Healthcare," *Int. J. Bio-*

Science Bio-Technology, vol. 5, no. 5, pp. 241–266, 2013.
[5] M. Bharati and M. Ramageri, "Data mining techniques and applications," 2010.
[6] S. G. Jacob and R. G. Ramani, "Evolving efficient clustering and classification patterns in lymphography data through data mining techniques," *Int. J. Soft Comput.*, vol. 3, no. 3, p. 119, 2012.
[7] A. Agrawal and S. Sharma, "Optimizing k-means for Scalability," *Int. J. Comput. Appl.*, vol. 120, no. 17, 2015.
[8] S. L. Ting, W. H. Ip, and A. H. C. Tsang, "Is Naive Bayes a good classifier for document classification?," *Int. J. Softw. Eng. Its Appl.*, vol. 5, no. 3, pp. 37–46, 2011.
[9] R. E. Putri, S. Suparti, and R. Rahmawati, "Perbandingan Metode Klasifikasi Naive Bayes dan k-Nearest Neighbor Pada Analisis Data Status Kerja Di Kabupaten Demak Tahun 2012," *J. gaussian*, vol. 3, no. 4, pp. 831–838, 2014.
[10] A. B. Hassanat, M. A. Abbadi, G. A. Altarawneh, and A. A. Alhasanat, "Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach," *arXiv Prepr. arXiv1409.0919*, 2014.
[11] M. Seetha, K. V. N. Sunitha, and G. M. Devi, "Performance assessment of neural network and k-nearest neighbour classification with random subwindows," *Int. J. Mach. Learn. Comput.*, vol. 2, no. 6, p. 844, 2012.
[12] H. S. Khamis, K. W. Cheruiyot, and S. Kimani, "Application of k-nearest neighbour classification in medical data mining," *Int. J. Inf. Commun. Technol. Res.*, vol. 4, no. 4, 2014.
[13] N. Bhatia, "Survey of nearest neighbor techniques," *arXiv Prepr. arXiv1007.0085*, 2010.
[14] S. B. Imandoust and M. Bolandraftar, "Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background," *Int. J. Eng. Res. Appl.*, vol. 3, no. 5, pp. 605–610, 2013.
[15] H. Parvin, H. Alizadeh, and B. Minaei-Bidgoli, "MKNN: Modified k-nearest neighbor," in *Proceedings of the World Congress on Engineering and Computer Science*, 2008, vol. 1.
[16] M. Karim and R. M. Rahman, "Decision tree and naive bayes algorithm for classification and generation of actionable knowledge for direct marketing," *J. Softw. Eng. Appl.*, vol. 6, no. 04, p. 196, 2013.
[17] Y. Yusra, "Perbandingan Klasifikasi Tugas Akhir Mahasiswa Jurusan Teknik Informatika Menggunakan Metode Naïve Bayes Classifier dan K-Nearest Neighbor," *J. Sains dan Teknol. Ind.*, vol. 14, no. 1, pp. 79–85, 2016.
[18] T. Efraim, E. A. Jay, T.-P. Liang, and R. V McCarthy, "Decision support systems and intelligent systems," *Up. Saddle River, NK Prentice Hall*, 2001.
[19] G. Kaur and S. Aggarwal, "Performance Analysis of Association Rule Mining Algorithms," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 8, pp. 856–858, 2013.
[20] S. Beniwal and J. Arora, "Classification and feature selection techniques in data mining," *Int. J. Eng. Res. Technol.*, vol. 1, no. 6, 2012.
[21] Y. A. Christobel and P. Sivaprakasam, "A New Classwise k Nearest Neighbor (CKNN) method for the classification of diabetes dataset," *Int. J. Eng. Adv. Technol.*, vol. 2, no. 3, pp. 200–396, 2013.
[22] T. N. Phyu, "Survey of classification techniques in data mining," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 2009, vol. 1, pp. 18–20.
[23] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.
[24] N. N. Faiza, "Prediksi Tingkat Keberhasilan Mahasiswa Tingkat I IPB dengan Metode k-Nearest Neighbor," 2009.
[25] V. Bijalwan, V. Kumar, P. Kumari, and J. Pascual, "KNN based machine learning approach for text and document mining," *Int. J. Database Theory Appl.*, vol. 7, no. 1, pp. 61–70, 2014.
[26] A. K. Santra and C. J. Christy, "Genetic algorithm and confusion matrix for document clustering," *Int. J. Comput. Sci. Issues*, vol. 9, no. 1, p. 322, 2012.